Published in final edited form as:

Commun Stat Theory Methods. 2009 September 1; 38(15): 2604–2619. doi:10.1080/03610920802585849.

A Two-Latent-Class Model for Smoking Cessation Data with Informative Dropouts

LI QIN¹, LISA A. WEISSFELD², CHANGYU SHEN³, and MICHELE D. LEVINE⁴

¹Center for Research on Health Care, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

²Department of Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

³Division of Biostatistics, School of Medicine, Indiana University, Indianapolis, Indiana, USA

⁴Department of Psychiatry, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Abstract

Non ignorable missing data is a common problem in longitudinal studies. Latent class models are attractive for simplifying the modeling of missing data when the data are subject to either a monotone or intermittent missing data pattern. In our study, we propose a new two-latent-class model for categorical data with informative dropouts, dividing the observed data into two latent classes; one class in which the outcomes are deterministic and a second one in which the outcomes can be modeled using logistic regression. In the model, the latent classes connect the longitudinal responses and the missingness process under the assumption of conditional independence. Parameters are estimated by the method of maximum likelihood estimation based on the above assumptions and the tetrachoric correlation between responses within the same subject. We compare the proposed method with the shared parameter model and the weighted GEE model using the areas under the ROC curves in the simulations and the application to the smoking cessation data set. The simulation results indicate that the proposed two-latent-class model performs well under different missing procedures. The application results show that our proposed method is better than the shared parameter model and the weighted GEE model.

Keywords

Area under ROC curve; Informative dropout; Latent class; Tetrachoric correlation

1. Introduction

Missing data is a common issue encountered in the analysis of longitudinal data. In the behavioral intervention setting, missed visits and/or losing to follow up can be extremely problematic. In this area, missed visits are assumed to be a result of failure of the intervention, sustained lack of interest in the study, or decreased desire to change the behavior. For smoking cessation and weight loss studies, these are common issues that must be dealt with at the data analysis phase. For example, Perkins et al. (2001) conducted a weight concern on smoking study. The purpose of this study is to determine if cognitive-behavioral therapy can reduce weight concern and increase the success of smoking cessation. The study includes 219 women

who were randomized to one of three groups: (i) behavioral weight control to prevent weight gain (weight control); (ii) cognitive-behavioral therapy to reduce concerns (CBT); or (iii) non specific social support (standard), which involved a discussion of weight. Participants were assessed for smoking abstinence, a binary measure, at 4-weeks postquit and 12-months postquit. However, the outcomes at the second time point for some women were missing due to drop-outs. The assumption in the smoking cessation literature is that these women were smoking so that all missing outcome values are set equal to zero (0 = smoking; 1 = not smoking) for the purposes of analyses. But it can introduce bias because not every woman who drops out is smoking.

In this article, we want to address the problem of informative missingness, in which the missing status depends on unknown outcome values. For this type of missingness, there are two main methods: selection models and pattern mixture models. In selection models, the joint distribution of outcomes and missingness is partitioned into the marginal distribution of outcome and the conditional distribution of missingness given outcomes. As an alternative to selection models, pattern mixture models work with the factorization of the joint distribution of outcomes and missingness into the marginal distribution of missingness and the conditional distribution of outcomes given missingness. Latent class models are another approach for informative missingness, and can be framed as a special case of a pattern mixture model, in which outcomes are divided into groups by the latent classes instead of by the missing patterns.

For latent variable models, variables are classified as ‘manifest’ when they can be directly observed and as ‘latent’ when they cannot be. A latent class model is a type of latent variable model with latent variables being categorical. Latent class models have been applied widely in the medical area for diagnoses (e.g., Garret and Zeger, 2000; Hadgu and Qu, 1998). Recently, latent class models are also used for dealing with missing data. Reboussin et al. (2002) proposed a latent class for multiple binary longitudinal outcomes subject to missing at random. The idea behind this method is to reweigh the binary outcomes by the inverse probability of being observed, which is an extension of Robins et al. (1995)’s weighted GEE approach. Roy (2003) proposed a latent dropout class model for continuous data with non ignorable dropouts. The ideas of latent dropout class models are based on the assumption that a small number of latent classes exist behind the sparse observed dropout times and that the probability of being in a given class is determined by the time of dropout.

Here, we propose a two-latent-class model for longitudinal binary response data with informative dropouts. In the proposed model, the observed data are divided into two latent classes: one of them is called ‘homogeneous’ class in which subjects have the same and deterministic outcomes (in the women’s smoking cessation study, we assume that the subjects in this class are always in smoking status) and a second one is called ‘heterogeneous’ class in which the subjects may have different outcomes and the outcomes can be modeled using the logistic regression. In the model, the latent variable is used as a mechanism to induce independence between the outcome and the missing status. Thus, in the proposed two-latent-class model, the drop-out process and response process are assumed to be independent given a latent class. Because these assumptions cannot be verified, we will assess the sensitivity by comparing the proposed model with other models such as the shared parameter model (Ten Have et al., 1998) and weighted GEE (Robins et al., 1995). Our ‘two-latent-class’ assumption cannot be used for all kinds of data, but this assumption is reasonable under some special cases, for example, in the smoking study, some subjects have much more vulnerability to smoking than the others and the outcomes are believed to be related to the genetic factors (Bergen and Caporaso, 1999).

2. Two-Latent-Class Model

The proposed two-latent-class model (TLCM) is based on the assumption that the binary responses are manifestation of latent classes. Here, we consider bivariate binary outcomes, which attempt to characterize the latent classes, and define $\mathbf{Y}_i = (Y_{i1} \ Y_{i2})'$, $i = 1, 2, \dots, n$, be a vector indicating whether i th subject quits smoking or not at different time points. We let $Y_{ij} = 1$ denote 'quit smoking' and $Y_{ij} = 0$ denote 'not quit smoking' for i th subject at time point j , $j = 1, 2$. Because of missingness, some subjects might only have Y_{i1} or have no outcomes at all. This kind of missing pattern is due to dropouts and may be related to unobserved outcomes, resulting in informative dropouts. Here we consider the settings where Y_{i1} is always observed and Y_{i2} could be observed or missing. Let R_i be the indicator denoting the missing status of subject i , where $R_i = 1$ if Y_{i2} is observed and 0 if Y_{i2} is missing.

Our method is motivated by the smoking cessation data, in which some subjects' outcomes are not affected by interventions or other factors. We define $\eta_i = (i = 1, \dots, n)$ to be i th subject's latent class, and $\eta_i = 1$ or 0. We assume that $\eta_i = 0$, if subject i is in a 'homogeneous' class, such that $Pr(\mathbf{Y}_i = (0, 0)' \mid \eta_i = 0) = 1$. From this assumption, in the class $\eta_i = 0$, the outcomes are $(0, 0)'$ only; in the class $\eta_i = 1$, the outcomes could be $(0, 0)'$, $(0, 1)'$, $(1, 0)'$ or $(1, 1)'$. In the smoking study, it can be explained as stubborn smokers, who cannot quit smoking at all. Furthermore, when $\eta_i = 1$ the subject i is considered to be in a 'heterogeneous' class and the outcomes are affected by the interventions and other factors.

Suppose for the subjects in the 'heterogeneous' class, the conditional outcome probability, $p_{ij} = Pr(Y_{ij} = 1 \mid \mathbf{x}_{ij}, \eta_i = 1)$, can be fit by the following logistic regression,

$$\text{logit}(p_{ij}) = \beta^T \mathbf{x}_{ij}$$

or

$$p_{ij} = \frac{\exp(\beta^T \mathbf{x}_{ij})}{1 + \exp(\beta^T \mathbf{x}_{ij})}, \quad (1)$$

where $\mathbf{x}_{ij}$ is a vector of covariates (possibly time-dependent) and β is a vector of regression coefficients. Note that p_{ij} depends on the same regression coefficients at different time points. (It is easy to extend this so that p_{ij} depends on different regression coefficients for different j .)

We let p_{i1} and p_{i2} denote the probabilities at time points 1 and 2 for the i th subject, and the joint probability $q_{ist} = Pr(Y_{i1} = s, Y_{i2} = t \mid \mathbf{x}_{ij}, \eta_i = 1)$, $s = 0, 1$, and $t = 0, 1$. Here, η_i is introduced as a latent trait that determines the two general categories of smoking habit, conditional on which the dropouts and outcomes are independent. So in our proposed model, the latent class is not assumed to explain the within-subject correlation between the outcomes. To calculate q_{i11} , q_{i10} , q_{i01} , and q_{i00} from p_{i1} and p_{i2} , we have to consider the correlation between Y_{i1} and Y_{i2} . Here we use the tetrachoric correlation, which is extended from probit marginals Ashford and Sowden (1970) to the logistic marginals by le Cessie and van Houwelingen (1994). The general idea is to obtain q_{i11} by using bivariate standard normal distributions and tetrachoric series,

$$\begin{aligned}
 q_{i11} &= \int_{-\infty}^{g_{i1}} \int_{-\infty}^{g_{i2}} f(t_1, t_2, \rho) dt_2 dt_1 \\
 &= p_{i1} p_{i2} + n(g_{i1}) n(g_{i2}) \sum_{k=0}^{\infty} \frac{1}{(k+1)!} H_k(g_{i1}) H_k(g_{i2}) \rho^{k+1},
 \end{aligned} \tag{2}$$

where $f(t_1, t_2, \rho)$ is the joint density function of the standardized bivariate normal distribution with correlation ρ , $n(u) = (2\pi)^{-1/2} \exp(-u^2/2)$ are the density function of the standard normal distribution, π is the correlation of Y_{i1} and Y_{i2} , $g_{ij} = \varphi^{-1}(p_{ij})$ with $\varphi^{-1}(\cdot)$ being the inverse of the standard normal cumulative distribution function, and

$$H_k(v) = \sum_{i=0}^{\lfloor k/2 \rfloor} \frac{k!}{i! (k-2i)!} (-1)^i 2^{-i} v^{k-2i} \tag{3}$$

are the Hermite polynomials, where $\lfloor k/2 \rfloor$ is the largest integer in the range of $\leq k/2$. After obtaining p_{i1} , p_{i2} , and q_{i11} , we can calculate q_{i10} , q_{i01} , and q_{i00} directly, for example, $q_{i10} = p_{i1} - q_{i11}$.

For the subjects in the homogeneous class, $Pr(Y_{ij} = 0 | \eta_i = 0) = 1$, that is, all the subjects in the homogeneous class have the outcome $Y_{ij} = 0$. In the study, we need know the probability of $\eta_i = 0$ or 1 conditional on the missing status. To do this, we left $\lambda_i = Pr(\eta_i = 1 | r_i)$ and fit the model

$$\text{logit}(\lambda_i) = \alpha_0 + \alpha_1 r_i, \tag{4}$$

where r_i is the observed value of R_i . For the probabilities of latent classes given the missingness, $Pr(\eta_i | r_i)$, we set $\mathbf{e} = (e_1, e_0)$, where $e_1 = Pr(\eta_i = 1 | R_i = 1)$ and $e_0 = Pr(\eta_i = 1 | R_i = 0)$. From the model above, we can obtain

$$e_0 = Pr(\eta_i = 1 | R_i = 0) = \frac{\exp(\alpha_0)}{1 + \exp(\alpha_0)}, \tag{5}$$

and

$$e_1 = Pr(\eta_i = 1 | R_i = 1) = \frac{\exp(\alpha_0 + \alpha_1)}{1 + \exp(\alpha_0 + \alpha_1)}. \tag{6}$$

Because we cannot observe the values of latent class η_i from the data, we will estimate e_0 and e_1 from the maximum likelihood estimation.

3. Maximum Likelihood Estimation

3.1. Likelihood

Letting $\mathbf{y}_i$ denote the vector of observed longitudinal binary responses for the i th subject, the likelihood for our proposed model is

$$\begin{aligned}
L(\mathbf{y}, r|\mathbf{x}) &= \prod_{i=1}^n L(\mathbf{y}_i|\mathbf{x}_i, r_i) L(r_i) \\
&= \prod_{i=1}^n [L(\mathbf{y}_i|\mathbf{x}_i, r_i, \eta_i=1) Pr(\eta_i=1|r_i) L(r_i) + L(\mathbf{y}_i|\mathbf{x}_i, r_i, \eta_i=0) Pr(\eta_i=0|r_i) L(r_i)] \\
&= \prod_{i=1}^n [L(\mathbf{y}_i|\mathbf{x}_i, \eta_i=1) Pr(\eta_i=1|r_i) L(r_i) + L(\mathbf{y}_i|\mathbf{x}_i, \eta_i=0) Pr(\eta_i=0|r_i) L(r_i)].
\end{aligned} \tag{7}$$

Here, we let $L(\mathbf{y}_i|\mathbf{x}_i, r_i, \eta_i=1) = L(\mathbf{y}_i|\mathbf{x}_i, \eta_i=1)$ and $L(\mathbf{y}_i|\mathbf{x}_i, r_i, \eta_i=0) = L(\mathbf{y}_i|\mathbf{x}_i, \eta_i=0)$, that is, given the latent class, η_i , the outcome, $\mathbf{Y}_i$, is independent of the missingness, R_i . This is an important assumption which reduces the mathematical complexity for estimation.

Based on the description above, the likelihood function including the parameters of interest can be written as

$$\begin{aligned}
L(\beta, \mathbf{e}, \rho; \mathbf{y}, r|\mathbf{x}) &= \prod_{i=1}^n [L(\beta, \rho, \mathbf{y}_i|\mathbf{x}_i, \eta_i=1) Pr(\eta_i=1|r_i) \\
&\quad + L(\beta, \rho, \mathbf{y}_i|\mathbf{x}_i, \eta_i=0) Pr(\eta_i=0|r_i)] L(r_i) \\
&= \prod_{i=1}^n [(Pr(\mathbf{y}_i|\mathbf{x}_i, \eta_i=1) e_1 + I(Y_{i1}=0, Y_{i2}=0) (1 - e_1)) I(R_i=1) \\
&\quad (Pr(\mathbf{y}_i|\mathbf{x}_i, \eta_i=1) e_0 + I(Y_{i1}=0) (1 - e_0)) I(R_i=0) L(r_i)],
\end{aligned} \tag{8}$$

where $I(\cdot)$ is the indicator function, and we assume that $Pr(y_{i1}, Y_{i2} \text{ is missing} | \mathbf{x}_{i1}, \eta_i = k) = Pr(y_{i1} | \mathbf{x}_{i1}, \eta_i = k)$, $k = 0, 1$, that is, missing value of Y_{i2} can be ignored after conditioning on η_i . $Pr(\mathbf{y}_i|\mathbf{x}_i, \eta_i=1)$ and $Pr(y_{i1} | \mathbf{x}_{i1}, \eta_i=1)$ can be obtained from Eqs. (1)–(3). Note that in Eq. (8), $L(r_i)$, the marginal distribution of R_i , does not concern with the parameters we are interested in, and thus can be ignored when maximizing the likelihood. The likelihood Eq. (8) indicates that if a subject has outcomes $Y_1 = 0$ and $Y_2 = 0$ or missing, then the subject could be in the homogeneous class or heterogeneous class.

3.2. Estimation

We let y_{ist} be the observed value of Y_{ist} , $Y_{ist} = 1$ if $Y_{i1} = s$ and $Y_{i2} = t$ for subject i , and $Y_{ist} = 0$ otherwise, $s, t = 0, 1$. Because $Pr(\mathbf{y}_i|\mathbf{x}_i, \eta_i=1) = q_{i11}^{y_{i11}} q_{i10}^{y_{i10}} q_{i01}^{y_{i01}} q_{i00}^{y_{i00}}$, $Pr(\mathbf{y}_i = (0, 0) | \mathbf{x}_i, \eta_i=0) = 1$, $Pr(y_{i1}|\mathbf{x}_{i1}, \eta_i=1) = p_{i1}^{y_{i1}} (1 - p_{i1})^{1-y_{i1}}$, and $Pr(y_{i1} = 0 | \mathbf{x}_{i1}, \eta_i=0) = 1$, the likelihood function (8) can be written as

$$L(\beta, \mathbf{e}, \rho; \mathbf{y}, r|\mathbf{x}) = \prod_{i=1}^n [q_{i11}^{y_{i11}} q_{i10}^{y_{i10}} q_{i01}^{y_{i01}} e_1^{1-y_{i00}} (q_{i00} e_1 + 1 - e_1)^{y_{i00}} I(R_i=1) + (p_{i1} e_0)^{y_{i1}} (1 - p_{i1} e_0)^{1-y_{i1}} I(R_i=0)] L(r_i). \tag{9}$$

Based on Eq. (9), the log-likelihood function is

$$\begin{aligned}
l(\beta, \mathbf{e}, \rho; \mathbf{y}, r|\mathbf{x}) &= \sum_{i=1}^n [(1 - y_{i00}) (y_{i11} \log q_{i11} + y_{i10} \log q_{i10} + y_{i01} \log q_{i01} + \log e_1) \\
&\quad + y_{i00} \log (q_{i00} e_1 + 1 - e_1)] I(R_i=1) \\
&\quad + \sum_{i=1}^n [y_{i1} (\log p_{i1} + \log e_0) + (1 - y_{i1}) \log (1 - p_{i1} e_0)] I(R_i=0).
\end{aligned} \tag{10}$$

We use the quasi-Newton method to obtain the estimates of β , ρ , e_1 , and e_0 by maximizing the marginal log-likelihood in Eq. (10). Initial values for β and ρ may be obtained from the standard GEE model (Liang and Zeger, 1986) by assuming that outcomes are missing completely at random (MCAR), that is, in GEE, the β and ρ are estimated using all available data set. The standard errors of the estimates are obtained from the inverse of the information matrix and substituting the maximum likelihood estimates (for the differentiation of log-likelihood with respect to β , ρ , e_1 , and e_0 , and the calculation of the information matrix; see Appendix). The *optim()* function in R software is applied to perform the above estimation procedures. In *optim()*, method 'L-BFGS-B' is used (Byrd et al., 1995) and each parameter is given a lower and upper bound besides the initial value. Based on $\hat{e}_1$, $\hat{e}_0$ and Eqs. (5), (6), we can obtain $\hat{\alpha}_0$ and $\hat{\alpha}_1$ directly.

The proposed latent class model can also be applied to intermittent missing data, and the likelihood function becomes

$$L(\beta, \mathbf{e}, \rho; \mathbf{y}, \mathbf{r} | \mathbf{x}) = \prod_{i=1}^n \left[\left(Pr(\mathbf{y}_i | \mathbf{x}_i, \eta_i=1) e_{11} + I(Y_{i1}=0, Y_{i2}=0) (1 - e_{11}) \right) I(\mathbf{R}_i=(1, 1)') \right. \\ \left. + \left(Pr(y_{i1} | \mathbf{x}_{i1}, \eta_i=1) e_{10} + I(Y_{i1}=0) (1 - e_{10}) \right) I(\mathbf{R}_i=(1, 0)') \right. \\ \left. + \left(Pr(y_{i2} | \mathbf{x}_{i2}, \eta_i=1) e_{01} + I(Y_{i2}=0) (1 - e_{01}) \right) I(\mathbf{R}_i=(0, 1)') \right] L(\mathbf{r}_i). \quad (11)$$

Here, $\mathbf{R}_i$ is a 2×1 vector, (R_{i1}, R_{i2}) , of indicator variables denoting the missing status of a subject i , where $R_{ij} = 1$ if Y_{ij} is observed and 0 if Y_{ij} is missing, where $j = 1, 2$; and $e_{st} = Pr(\eta_i = 1 | R_{i1} = s, R_{i2} = t)$, $s = 0, 1$, and $t = 0, 1$. It is fairly straightforward to apply the extension to the actual data sets by including the intermittent missing-data patterns. For instance, one can use a logistic regression model to treat the latent classes as the outcomes and the missing-data patterns as covariates. But *optim()* in R may not be applied to them because the likelihood for the complicated cases tends to be relatively flat, which can lead to some maximization algorithms having difficulty in converging. We may use the modified EM algorithm proposed by Lin et al. (2000) to obtain the maximum likelihood estimator (MLE) for them.

4. Shared Parameter Model and Weighted GEE Model

Ten Have et al. (1998) developed a shared parameter model (SPM) with a logistic link for longitudinal binary response data to accommodate informative dropouts. The model includes two components: observed longitudinal components and dropout components. These two parts share random effects parameters and they are independent after conditioning on the random effects structure. The shared random effects here are assumed to be continuous and they count for the correlation within subjects and the correlation between the outcomes and the missing status. It is one of the differences between the shared parameter model and the proposed two-latent-class model. In our proposed method, the correlation within subjects is considered separately.

Robins et al. (1995) proposed weighted GEE model (WGEE) in which the parameter estimates are consistent when the responses are missing at random (MAR). In weighted GEE model, weights equal to the inverse probability of 'being missing' at time of attrition are added to the standard GEE model.

5. Simulation Results

We perform a simulation study comparing the proposed method with the shared parameter model and the weighted GEE model. We generate data by considering two aspects: the logistic model structure for outcomes and the missing structure.

For the outcomes, we consider the case of a binary response measured at two time points, $\mathbf{Y}_i = (Y_{i1}, Y_{i2})'$ with correlation ρ . To generate data for this scenario, we first generate the continuous variables $\mathbf{Z}_i = (Z_{i1}, Z_{i2})'$, which are from a bivariate standard normal distribution with correlation δ . Then we let $Y_{ij} = 1$ if $Z_{ij} \leq \Phi^{-1}(p_{ij})$; $Y_{ij} = 0$, if $Z_{ij} > \Phi^{-1}(p_{ij})$, and Φ^{-1} is the inverse of the cumulative distribution function for a standard normal variable. Here, p_{ij} are the marginal probabilities, $p_{ij} = Pr(Y_{ij} = 1 | \mathbf{x}_{ij}) = E(Y_{ij} | \mathbf{x}_{ij})$, and are obtained from $\text{logit}(p_{ij}) = \beta x_{1i} + \beta x_{2ij}$, with x_{1i} is from a standard normal distribution, $x_{2i1} = 0.1$ and $x_{2i2} = 0.2$; $\beta = 2$. For the purpose of simulation, the correlation, δ , is set to be 0.5. According to Emrich and Piedmonte (1991), the correlation between Y_{i1} and Y_{i2} is given by, $\rho = [\Phi(\Phi^{-1}(p_{i1}), \Phi^{-1}(p_{i2}), \delta) - p_{i1}p_{i2}] / [p_{i1}(1 - p_{i1})p_{i2}(1 - p_{i2})]^{1/2}$, where Φ is the cumulative distribution function for a standard bivariate normal variable with correlation coefficient ρ , and Φ^{-1} is the inverse of the cumulative distribution function for a standard normal variable.

For the missing structure, we assume a monotone missing data pattern with the binary response at the first time point completely observed. Three missing procedures are considered for the response at time point 2: (i) MAR missing procedure, in which

$$\text{logit} [Pr(R_i=0|x_{1i})] = -0.5 - 0.5x_{1i}.$$

Because the missingness depends on the observed data, this missing procedure is missing at random (MAR). (ii) SPM missing procedure. For this missing procedure, note that the $p_{ij} = Pr(Y_{ij} = 1 | \mathbf{x}_{ij}, \tau)$ are obtained from

$$\text{logit}(p_{ij}) = \beta x_{1i} + \beta x_{2ij} + \sigma \tau,$$

where τ is a normal variable with mean = 0 and variance $\sigma^2 = 1$ or $\sigma^2 = 5^2$ that is, the correlation between Y_{i1} and Y_{i2} is counted from the random effect τ . For the missing process,

$$\text{logit} [Pr(R_i=0|\tau)] = -0.5 + \sigma \tau.$$

According to Ten Have et al. (1998), when $\sigma = 1$, the dependency of dropouts on the random effect variable is moderate; when $\sigma = 5$, the dependency is strong. We want to see if the proposed two-latent-class model can handle different levels of dependency. (iii) TLCM missing procedure, in which we let $Pr(R_i = 0) = 0.25$. In the data generation procedure, for MAR and SPM missing procedures, we first obtain full data set, then according to the missing structure to delete some Y_{i2} . But in TLCM missing procedure, it is different: after obtaining missing patterns, we define the latent classes in which we let $e_1 = Pr(\eta_i = 0 | Y_{i2} \text{ is observed})$, $e_0 = Pr(\eta_i = 0 | Y_{i2} \text{ is missing})$, and we assume that Y_{i1} is always observed. We generate the outcome structure in each latent class and according to the value of R_i to delete some Y_{i2} . For the latent classes, η_i , $\eta_i = 0$ denotes the homogeneous case where $Pr(Y_{ij} = 0 | \eta_i = 0) = 1$; and $\eta_i = 1$ denotes the heterogeneous case where $Pr(Y_{ij} = 1 | \mathbf{x}_{ij}, \eta_i = 1) = \exp(\beta x_{1i} + \beta x_{2ij}) / \{1 + \exp(\beta x_{1i} + \beta x_{2ij})\}$. For the simulation, we considered $e_1 = 0.95$ and $e_0 = 0.80$ or $e_1 = 0.70$ and $e_0 = 0.55$ to examine the impacts affected by the different latent classes of the data on the model fit.

For the entire simulation study, we consider the three types of missing procedures and sample sizes of 200 with 1,000 replications. Because in our proposed model, outcomes are divided into two latent classes and in one of them, outcomes cannot be modeled, so we cannot obtain marginal estimates of parameters. The comparisons between our proposed model and the other models will be performed using the area under the Receiver Operating Characteristic (ROC) curve. We use trapezoidal rule to calculate the area under the curve (AUC), which is a nonparametric method based on constructing trapezoids under the curve as an approximation of area. The summary measures for each model are mean AUC and standard error of mean AUC over 1,000 replications, which are based on estimated outcomes Y_{i2} by each model and the complete Y_{i2} in simulation data before the deletion as a gold standard. Table 1 presents the simulation results. From the results of the table, we see that the two-latent-class model and the shared parameter model performed optimally under its own missing procedure. When the missing procedure is TLCDM with $e_1 = 0.95$ and $e_0 = 0.80$, the two-latent-class model is the best one (mean AUC = 0.888), but the shared parameter model and the weighted GEE model performs well too, and both of them have mean AUC around 0.82. But when $e_1 = 0.70$ and $e_0 = 0.55$, that is, more subjects belong to the ‘homogeneous’ latent class in which outcomes cannot be modeled, the two-latent-class model keeps performing well (mean AUC = 0.888), but mean AUCs under the shared parameter model and the weighted GEE methods drop to around 0.75. When the missing procedure is SPM with $\sigma = 1$, that is, the dependency of dropout on the random effect variable is moderate, all methods perform well (for mean AUC, shared parameter model is 0.880; two-latent-class model is 0.858; weighted GEE model is 0.835). But when $\sigma = 5$, that is, the dependency of dropout on the random effect variable is strong, although the performance of the shared parameter model is the best one (mean AUC = 0.963), the two-latent-class model’s mean AUC is 0.863, which is good too. In this case, the weighted GEE model performs bad and the mean AUC is 0.664. When the missing procedure is MAR, in which missingness depends on the observed data, all methods performs well with the two-latent-class model having largest mean AUC (0.892). Because of the relatively small standard errors of the mean AUCs in the results, the above differences in mean AUCs among the three models are statistically significant (p -value < 0.001).

In the comparisons among the three methods, the overall conclusions are that the two-latent-class model performs well under all missing procedures; the shared parameter model performs well in most cases except under TLCDM missing procedure with small e_1 and e_0 ; the weighted GEE model performs well when the missing procedure is MAR or missingness is ‘weak’ informative.

6. Application to the Smoking Cessation Study

In this section, we illustrate the proposed two-latent-class model, the shared parameter model and the weighted GEE model using an example from the women’s smoking cessation study (Perkins et al., 2001). This is a longitudinal study designed to assess the effect of weight concern on smoking cessation for women. At enrollment, 219 women met the eligibility criteria. All of the participants were randomly divided into three groups: (i) behavioral weight control to prevent weight gain (weight control); (ii) cognitive-behavioral therapy to reduce weight concerns (CBT); or (iii) non specific social support (standard), which was a control group and involved no discussion of weight. Each of the three interventions consisted of ten 90-minute sessions provided over 7 weeks, with two sessions per week during the first 3 weeks and 1 session per week over the next 4 weeks. Participants were instructed to quit smoking after the fourth session. Follow-up sessions were scheduled at 3, 6, and 12 months postquit for assessment purposes; no treatment was provided in these periods. In this trial, the repeated binary responses of interest are whether the participants are in continuous abstinence or not (1 = yes, 0 = no). Here, continuous abstinence was defined as no relapse since the quit day and relapse were defined as self-report of 7 consecutive days of any smoking at all or an expired-

air carbon monoxide (CO) greater than 8 ppm, as widely recommended (Ossip-Klein et al., 1986).

In this study, we focus on the outcomes at two time points, 4-week postquit (Y_1) and 12-month postquit (Y_2). The 57 women who had missing data at 4-week postquit (also missing at 12-month postquit) were removed from all analyses, leaving 162 subjects (116 subjects have no missing data; 46 subjects were observed at 4-week postquit and missing at 12-month postquit). To identify significant covariates related to outcomes, we carried out a preliminary analysis by using standard GEE (Liang and Zeger, 1986). The results showed that the following variables should be included in the models: group (*weight control*, *CBT*) ('standard' as a control group), time (t) and age at first cigarette (*age*). We also fit a generalized additive model (GAM) for the outcomes, Y_1 , with the covariate, 'age at first cigarette', that is, $\text{logit}(E(Y_1)) = s_0 + s_1(\text{age})$, where $s_i(\cdot)$, $i = 0$ and 1 , are smooth functions. Figure 1 shows the plot of smooth function in GAM with x -axis presenting the variable of 'age at first cigarette' and y -axis presenting smooth function of the variable. From the result of GAM, we add the square term for 'age at first cigarette', in addition to the linear term.

A summary of outcomes is in Table 2. It shows that missingness at 12-month follow-up is 28.40%. The abstinent rate for subjects without missingness at 12-month follow-up is $53/(53 + 63) = 45.68\%$, which is much less than the abstinent rate (72.84%) at the 4-week postquit. It also shows that the dropout might be related to the unobserved second-time outcomes. Based on these results it is reasonable to consider nonignorable missingness.

In Table 3, we present the parameter estimates, standard errors and Z-values calculated from them for these three models. These parameter estimates are common fixed effects under these models. From the results, we can see that 'time' is significant in the models with the decreasing abstinent rate over time. All of the analyses show that the 'CBT' group has a larger abstinent rate compared with the 'standard' group. The results obtained from the weighted GEE model also show that 'age at first cigarette' is significant in the square term, while this factor is not significant in the other models. In the two-latent-class model, we obtain the correlation within the subject, $\rho = 0.347$ (SE = 0.189, $Z = 1.84$). Table 4 gives the estimates for e_1 and e_0 . It indicates that subjects with missing outcomes have a greater probability ($1 - 0.780 = 0.220$) of being in the special status, that is, they are stubborn smokers, than the subjects without missing outcomes ($1 - 0.817 = 0.183$). According to the estimates and the standard errors in Table 4, we can compare the difference between e_1 and e_0 and t -test statistics is 4.35 with 160 degrees of freedom (p -value < 0.001), so the subjects dropping out the study have significantly more probability of being in the 'homogeneous' latent class, that is, being smoking. From the $\hat{e}_1$ and $\hat{e}_0$, we obtain $\hat{\alpha}_1 = 0.010$ and $\hat{\alpha}_0 = 0.550$.

We compare the performance of the proposed two-latent-class model, the shared parameter model and the weighted GEE model in terms of the empirical ROC curve and its area under the curve (AUC). The calculations are based on estimated outcomes at three months by each model and the observed data as a gold standard. We summarize the estimated ROC curves for these three methods in Fig. 2, and the corresponding areas under the ROC curves with standard errors for the two-latent-class model, the shared parameter model and the weighted GEE model are 0.865 ± 0.030 , 0.623 ± 0.046 , and 0.629 ± 0.046 , respectively. The computation of AUC and the standard error of AUC is carried out using trapezoidal rule and the variance of the Wilcoxon statistic (Hanley and McNeil, 1982). According to the method given by Hanley and McNeil (1983), the critical ratio between the areas estimated by the two-latent-class model and the shared parameter model is 4.92 (p -value < 0.001). The critical ratio between the areas estimated by the two-latent-class model and the weighted GEE model is 4.77 (p -value < 0.001). It appears that the proposed two-latent-class model gives significantly better fit to the data than the other methods.

7. Diagnostics for the Two-Latent-Class Model

Pregibon (1981) extended regression diagnostics for logistic regression. Based on his article, the component of the χ^2 goodness-of-fit statistic for subject i at time point j , is

$$\chi_{ij} = \frac{Y_{ij} - \hat{\alpha}_{ij}}{\sqrt{\hat{\alpha}_{ij}(1 - \hat{\alpha}_{ij})}},$$

where $\alpha_{ij} = E(Y_{ij})$ is the marginal probability and $\hat{\alpha}_{ij}$ is the estimates of α_{ij} based on the model. In the two-latent-class model, the marginal estimate, $\hat{\alpha}_{ij}$, is calculated as

$$\begin{aligned}\hat{\alpha}_{ij} &= Pr(Y_{ij}=1|\eta_i=1) Pr(\eta_i=1) + Pr(Y_{ij}=1|\eta_i=0) Pr(\eta_i=0) \\ &= \hat{p}_{ij} Pr(\eta_i=1),\end{aligned}$$

where $\hat{p}_{ij}$ is estimated under Eq. (1), and

$$\begin{aligned}Pr(\eta_i=1) &= Pr(\eta_i=1|R_i=1) Pr(R_i=1) + Pr(\eta_i=1|R_i=0) Pr(R_i=0) \\ &= \hat{e}_0 Pr(R_i=1) + \hat{e}_1 Pr(R_i=0).\end{aligned}$$

Figure 3 are index plots of the χ^2 components for the two-latent-class model at each time point. Asymptotic arguments concerning the distribution of the χ^2 goodness-of-fit statistic suggest that χ_{ij} should be of order of magnitude $N(0, 1)$. Observations with values outside ± 2 are not well accounted for by the fitted model (Kay and Little, 1986). From the figure, we can see that there are very few poorly fit subjects in the two-latent-class model.

Table 5 lists observations $i = 114, 127, 135, 143$, and 151 , which are poorly fit by the two-latent-class model. The common characteristics of these subjects are that they were 11–13 years old at first cigarette, they belonged to nonspecific social support treatment group, and they quit smoking at 12-month postquit. Based on the results in Table 3, the estimated probabilities of being quitting smoking for these subjects are very small. It explains that why these five subjects are outliers.

8. Discussion

In this article, we develop a two-latent-class model for categorical responses with non ignorable dropouts, in which subjects are divided into two latent classes and one of them is deterministic. The dropout time is related to the latent class, whose probability is estimated by the maximum likelihood estimation. For the relationship within subjects, we use the tetrachoric correlation (le Cessie and van Houwelingen, 1994) for the estimation.

The simulation results and the application to the smoking cessation data provide support for the use of the proposed two-latent-class model when compared to the shared parameter model and weighted GEE model. The results of the simulation indicate that the proposed model generally performs well in various missing procedures. Although the shared parameter model is also good in most cases, when data is heterogeneous with outcomes in different strata, the two-latent-class model is a better choice. In the results of the application to the smoking cessation data, the parameter estimates of the proposed model for the subjects in the heterogeneous class are similar to the shared parameter model for all the subjects. But the proposed model consider the data set divided into different latent classes, which results in better

AUC, that is, the proposed model fits the data in a better way. But because we only consider the two time point paradigm in the simulation and application, the above conclusion is fairly limited.

In the two-latent-class model, the latent variable is used to induce conditional independence between the outcome and missing status so that standard likelihood techniques can be used to derive the estimators. While the two-latent-class model can be considered as a type of pattern mixture model, the difference between them is that in the two-latent-class model, ‘pattern’ is not based on missing patterns but on the strata behind the data set, for example, in the smoking cessation data, we assume that there exist two classes based on the smoking stubborn. Here, we focused on monotone missingness, but the method can also be used for intermittent missing data in the same way.

One of the advantages of our proposed method is that its calculation is easy because there are only two latent classes and in one of them the outcomes are deterministic. The other advantage is according to the simulation results, the two-latent-class model can be used under various missing procedures. Our method can be easily extended for the situation of multivariate outcomes, that is, the total time points, J , is larger than 2. We know that as J increases, the calculation of the tetrachoric correlation becomes complicated. One of the solutions to this problem is that we can use the product of all pairwise likelihood with a subject instead of the true contribution of a subject to the likelihood (le Cessie and van Houwelingen, 1994). For example, for i th subject, its contribution to the log-likelihood is

$$l_i = \frac{1}{J-1} \sum_{j=1}^J \sum_{j'=1}^j l_{jj'}, \quad (12)$$

where $l_{jj'}$ is the pairwise likelihood of $(Y_j, Y_{j'})$ and the calculation of $l_{jj'}$ refers to Eq. (10). When both Y_j and $Y_{j'}$ are missing, we let $l_{jj'} = 0$. The multivariate outcome situation will be considered as one of our future works besides extending the method to other distributions, for example, ordinal responses.

Acknowledgments

This research was supported by National Institute on Drug Abuse R01 DA04174.

Appendix

Differentiation of the log-likelihood function (10) with respect to β , ρ , e_1 , and e_0 yields

$$\begin{aligned} \frac{\partial l}{\partial \beta} &= \sum_{i=1}^n \left[\frac{y_{i11}}{q_{i11}} \frac{\partial q_{i11}}{\partial \beta} + \frac{y_{i10}}{q_{i10}} \frac{\partial q_{i01}}{\partial \beta} + \left(\frac{y_{i00}}{q_{i00} + 1 - e_1} \frac{\partial q_{i00}}{\partial \beta} \right) \right] I(R_i=1) \\ &\quad + \sum_{i=1}^n \left(\frac{y_{i1}}{p_{i1}} \frac{\partial p_{i1}}{\partial \beta} - \frac{1-y_{i1}}{1-p_{i1}e_0} \frac{\partial p_{i1}}{\partial \beta} \right) I(R_i=0), \\ \frac{\partial l}{\partial \rho} &= \sum_{i=1}^n \left(\frac{y_{i11}}{q_{i11}} - \frac{y_{i10}}{q_{i10}} - \frac{y_{i01}}{q_{i01}} + \frac{y_{i00}e_1}{q_{i00}e_1 + 1 - e_1} \right) \frac{\partial q_{i11}}{\partial \rho} I(R_i=1), \\ \frac{\partial l}{\partial e_1} &= \sum_{i=1}^n \left(\frac{1-y_{i00}}{e_1} - \frac{y_{i00}(q_{i00}-1)}{q_{i00}e_1 + 1 - e_1} \right) I(R_i=1), \\ \frac{\partial l}{\partial e_0} &= \sum_{i=1}^n \left(\frac{y_{i1}}{e_0} - \frac{(1-y_{i1})p_{i1}}{1-p_{i1}e_0} \right) I(R_i=0), \end{aligned}$$

Here, we have used that $\partial q_{i10}/\partial \rho = \partial(p_{i1} - q_{i11})/\partial \rho = -\partial q_{i11}/\partial \rho$, etc. To estimate $\frac{\partial q_{i11}}{\partial \beta}$ and $\frac{\partial q_{i11}}{\partial \rho}$, from Eq.(2) $q_{i11} = \int_{-\infty}^{g_{i1}} \int_{-\infty}^{g_{i2}} f(t_1, t_2, \rho) dt_2 dt_1$, we have

$$\frac{\partial q_{i11}}{\partial \beta} = \Phi \left\{ \frac{g_{i2} - \rho g_{i1}}{\sqrt{(1-\rho^2)}} \right\} \frac{\partial p_{i1}}{\partial \beta} + \Phi \left\{ \frac{g_{i1} - \rho g_{i2}}{\sqrt{(1-\rho^2)}} \right\} \frac{\partial p_{i2}}{\partial \beta}$$

and

$$\frac{\partial q_{i11}}{\partial \rho} = f(g_{i1}, g_{i2}, \rho),$$

where Φ is the standard normal cumulative distribution function.

The elements of the observed information matrix are obtained by taking the second-order derivatives on the log-likelihood function (10),

$$\begin{aligned} \frac{\partial^2 l}{\partial \beta^2} = & - \sum_{i=1}^n \left\{ \frac{y_{i11}}{q_{i11}^2} \left(\frac{\partial q_{i11}}{\partial \beta} \right) \left(\frac{\partial p_{i11}}{\partial \beta} \right)' + \frac{y_{i10}}{q_{i10}^2} \left(\frac{\partial q_{i10}}{\partial \beta} \right) \left(\frac{\partial p_{i10}}{\partial \beta} \right)' + \frac{y_{i01}}{q_{i01}^2} \left(\frac{\partial q_{i01}}{\partial \beta} \right) \left(\frac{\partial p_{i01}}{\partial \beta} \right)' \right. \\ & \left. + \frac{y_{i00} e_1}{(q_{i00} + e_1 + 1 - e_1)^2} \frac{y_{i00}}{q_{i00}^2} \left(\frac{\partial q_{i00}}{\partial \beta} \right) \left(\frac{\partial p_{i00}}{\partial \beta} \right)' \right\} \\ & - \sum_{i=1}^n \left\{ \frac{y_{i1}}{p_{i1}^2} \left(\frac{\partial p_{i1}}{\partial \beta} \right) \left(\frac{\partial p_{i1}}{\partial \beta} \right)' + \frac{(1-y_{i1}) e_0}{(1-p_{i1} e_0)^2} \left(\frac{\partial p_{i1}}{\partial \beta} \right) \left(\frac{\partial p_{i1}}{\partial \beta} \right)' \right\} I(R_i=0), \end{aligned}$$

References

- Ashford JR, Sowden RR. Multivariate probit analysis. *Biometrics* 1970;26:535–546. [PubMed: 5480663]
- Bergen AW, Caporaso N. Cigarette smoking. *J. Nat. Cancer Instit* 1999;16:1365–1375.
- Byrd RH, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained optimization. *SIAM J. Scientific Comput* 1995;16:1190–1208.
- Emrich LJ, Piedmonte MR. A method for generating high-dimensional multivariate binary variates. *Amer. Statistician* 1991;45:302–304.
- Garret ES, Zeger SL. Latent class model diagnosis. *Biometrics* 2000;56:1055–1067. [PubMed: 11129461]
- Hadgu A, Qu Y. A biomedical application of latent class models with random effects. *Appl. Statist* 1998;47:603–616.
- Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29–36. [PubMed: 7063747]
- Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983;148:839–843. [PubMed: 6878708]
- Kay R, Little S. Assessing the fit of the logistic model: a case study of children with the haemolytic uraemic syndrome. *Appl. Statist* 1986;35:16–30.
- le Cessie S, van Houwelingen JC. Logistic regression for correlated binary data. *Appl. Statist* 1994;43:95–108.
- Liang K, Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika* 1986;73:13–22.
- Lin H, McCulloch CE, Turnbull BW, Slate EH, Clark LC. A latent class mixed model for analyzing biomarker trajectories with irregularly scheduled observations. *Statist. Med* 2000;16:1303–1318.

- Ossip-Klein DJ, Bigelow G, Parker SR, Curry S, Hall S, Kirkland S. Classification and assessment of smoking behavior. *Health Psychol* 1986;5(Suppl.):3–11. [PubMed: 3582323]
- Perkins KA, Marcus MD, Levine MD, D'Amico D, Miller A, Broge M, Ashcom J. Cognitive-behavioral therapy to reduce weight concerns improves smoking cessation outcome in weight-concerned women. *J. Consult. Clin. Psycho* 2001;69:604–613.
- Pregibon. Logistic regression diagnostics. *Ann. Statist* 1981;9:705–724.
- Reboussin BA, Miller ME, Lohman KK, Ten Have TR. Latent class models for longitudinal studies of the elderly with data missing at random. *Appl. Statist* 2002;51:69–90.
- Robins JM, Rotnetzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statis. Assoc* 1995;90:106–121.
- Roy J. Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics* 2003;59:829–836. [PubMed: 14969461]
- Ten Have TR, Kunselman AR, Pulkstenis EP, Landis JR. Mixed effects logistic regression models for longitudinal binary response data with informative dropout. *Biometrics* 1998;54:367–383. [PubMed: 9544529]

Figure 1.

Plots of smooth functions in a generalized additive model for Y_1 with 'age at first cigarette (age)' as covariate for the smoking cessation study. (The dashed lines indicate plus and minus two pointwise standard deviations; the number in the y-label gives the effective degree of freedom for the smooth function estimate.)

Figure 2.

Estimated ROC curves for the smoking cessation data based on the two-latent-class model, the shared parameter model, and the weighted GEE model.

Figure 3.
Index plot for the component χ_i of the χ^2 goodness-of-fit statistic.

Table 1

Simulation results: mean AUC and standard error for the three models under different missing procedures

Model	Statistic	MAR	Missing procedure			
			SPM		TLCM	
			$\sigma = 1$	$\sigma = 5$	$e_1 = 0.95, e_0 = 0.80$	$e_1 = 0.70, e_0 = 0.55$
Weighted GEE model	AUC	0.858	0.835	0.664	0.819	0.748
	SE	0.002	0.002	0.004	0.003	0.003
Shared parameter model	AUC	0.862	0.879	0.963	0.829	0.758
	SE	0.002	0.002	0.001	0.003	0.004
Two-latent-class model	AUC	0.892	0.858	0.868	0.888	0.888
	SE	0.002	0.002	0.003	0.002	0.002

Table 2

Summary of outcomes for the smoking cessation study

		12-month postquit			Total
		y ₂ is missing	y ₂ = 0	y ₂ = 1	
4-week postquit	y ₁ = 0	13	26	5	44(27.16%)
	y ₁ = 1	33	37	48	118(72.84%)
	Total	46(28.40%)	63(38.89%)	53(32.72%)	162

Table 3

Marginal parameter estimates, estimated standard errors, and Z-values for the smoking cessation study (Modeling $\Pr(\text{abstinent})$)

Par.	WGEE			SPM			TLCM*		
	Est.	SE	Z	Est.	SE	Z	Est.	SE	Z
β_0	3.323	1.946	1.71	1.630	2.573	0.63	1.618	2.573	0.63
β_{WC}	0.483	0.361	1.34	0.770	0.462	1.67	0.756	0.462	1.64
β_{CBT}	0.859	0.396	2.17	1.046	0.512	2.05	1.145	0.512	2.24
β_t	-0.117	0.021	-5.63	-0.120	0.030	-4.02	-0.156	0.030	-5.22
β_{age}	-0.494	0.271	-1.82	-0.380	0.369	-1.03	-0.380	0.369	-1.03
β_{age^2}	0.021	0.009	2.21	0.020	0.013	1.49	0.023	0.013	1.74

* Estimates are for the subjects who belong to $\eta = 0$.

Table 4

Estimates, estimated standard errors, and Z-values for the latent classes, e_1 and e_0 , under the proposed method for the smoking cessation study (Z-value is for testing $H_0 : e_1 = 1$ or $H_0 : e_0 = 1$)

	Est.	SE	Z
e_1	0.817	0.032	5.64
e_0	0.780	0.054	4.08

Table 5

Observations poorly fit by the two-latent-class model

Observations (<i>i</i>)	Y_2	Group	Age at first cigarette (Years)
114	1	Standard	13
127	1	Standard	12
135	1	Standard	13
143	1	Standard	13
151	1	Standard	11